This listing of claims will replace all prior versions, and listings, of claims in the application:

Claim 1 (currently amended): A system for providing capitalization correction for unstructured excerpts, comprising:

a preprocessor to tokenize an excerpt of unstructured content into a set of words; and

a capitalizer to analyze the set of words for correct capitalization, comprising:

an evaluator to evaluate individual characters constituting at least one such word in the set of words; and

a filter to skip the at least one such word if determined to be of a predefined type such that the capitalizer does not capitalize the at least one such word.

Claim 2 (currently amended): A system according to Claim 1, further comprising:

a document title capitalizer to provide one or more of the words with an initial letter in uppercase and each remaining letter in lowercase.

Claim 3 (original): A system according to Claim 1, further comprising:

a sentence capitalizer to provide only an initial such word with an initial letter in uppercase and each remaining letter in lowercase.

Claim 4 (currently amended): A system according to Claim 1, ~~further comprising:~~

~~a word analyzer to skip at least one of each such~~ wherein the predefined type is one of (A) a word comprising a number, ~~each such~~ (B) a word including no vowels, and ~~each such~~ (C) a word not occurring at a start of a phrase and constituting at least one of an article, conjunction, preposition.

Claim 5 (original): A system according to Claim 1, further comprising:

a lexicon comprising one or more reference words with at least one reference word defining a form of capitalization for the reference word;

a matcher to match the at least one such word against the reference words, the evaluator skipping each such word if a matching reference word is found.

Claim 6 (currently amended): A system according to Claim 1, further comprising:

a proper noun capitalizer to provide each of the individual letters in each such word comprising a noun with no vowels in uppercase.

Claim 7 (currently amended): A system according to Claim 1, ~~further comprising:~~

~~a tokenizer to tokenize~~ wherein the preprocessor tokenizes the excerpt into the one or more words and one or more punctuation marks.

Claim 8 (currently amended): A method for providing capitalization correction for unstructured excerpts, comprising:

tokenizing an excerpt of unstructured content into a set of words; and

analyzing the set of words for correct capitalization, comprising:

evaluating individual characters constituting at least one such word in the set of words; and

skipping the at least one such word if determined to be of a predefined type such that the capitalizer does not capitalize the at least one such word.

Claim 9 (original): A method according to Claim 8, further comprising:

providing one or more of the words with an initial letter in uppercase and each remaining letter in lowercase.

Claim 10 (original): A method according to Claim 8, further comprising:

providing only an initial such word with an initial letter in uppercase and each remaining letter in lowercase.

Claim 11 (currently amended): A method according to Claim 8, ~~further comprising:~~

~~skipping at least one of each such~~ wherein the predefined type is one of (A) a word comprising a

number, ~~each such~~ (B) a word including no vowels, and ~~each such~~ (C) a word not occurring at a start of a phrase and constituting at least one of an article, conjunction, preposition.

Claim 12 (original): A method according to Claim 8, further comprising:

maintaining a lexicon comprising one or more reference words with at least one reference word defining a form of capitalization for the reference word;

matching the at least one such word against the reference words; and

skipping each such word if a matching reference word is found.

Claim 13 (currently amended): A method according to Claim 8, further comprising:

providing each of the individual letters in each such word comprising a noun with no vowels in uppercase.

Claim 14 (currently amended): A method according to Claim 8~~, further comprising:~~ wherein the act of tokenizing includes tokenizing the excerpt into the one or more words and one or more punctuation marks.

Claim 15 (canceled)

Claim 16 (currently amended): An apparatus for providing capitalization correction for unstructured excerpts, comprising:

means for tokenizing an excerpt of unstructured content into a set of words; and

means for analyzing the set of words for correct capitalization, comprising:

means for evaluating individual characters constituting at least one such word in the set of words; and

means for skipping the at least one such word if determined to be of a predefined type such that the capitalizer does not capitalize the at least one such word.

Claim 17 (currently amended): A system for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

a ripper assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

an aggregator aggregating each word set, comprising:

an analyzer identifying at least one word set comprising significant statistics; and

a non-standard capitalization selector selecting at least ~~one~~ two such ~~variation~~ variations within the identified word set having a non-standard capitalization, and adding the at least ~~one~~ two such ~~variation~~ variations to the lexicon.

Claim 18 (original): A system according to Claim 17, further comprising:

a tokenizer tokenizing the excerpt into the one or more words and one or more punctuation marks.

Claim 19 (original): A system according to Claim 18, wherein hyphenated words are split into a plurality of the words.

Claim 20 (original): A system according to Claim 17, wherein at least one variation appearing at the start of a sentence is skipped.

Claim 21 (original): A system according to Claim 20, wherein the non-standard capitalization comprises the at least one variation occurring in an excerpt having fewer than half of individual letters provided in uppercase.

Claim 22 (currently amended): A system according to Claim 17, further comprising:

a normalizer normalizing a plurality of the words extracted relative to a source of the unstructured excerpt ~~structured Web content~~.

Claim 23 (currently amended): A system according to Claim 17, wherein the set comprising significant statistics comprises only non-standard capitalization variations having at least four occurrences of at least one such variation within a word set.

Claim 24 (original): A system according to Claim 17, wherein the non-standard capitalization comprises the at least one variation having any individual letter other than the first individual letter provided in uppercase.

Claim 25 (currently amended): A system according to Claim 17, further comprising:

a standard capitalization selector selecting at least ~~one~~ two such ~~variation~~ variations within the identified word set having a standard capitalization, and adding the at least ~~one~~ two such ~~variation~~ variations to the lexicon.

Claim 26 (currently amended): A system according to Claim 17, further comprising:

a validator applying implicit rules for capitalization, and skipping each of the at least ~~one~~ ~~variation~~ two variations subject to at least one such implicit rule.

Claim 27 (currently amended): A system according to Claim 26, wherein the implicit rules comprise skipping each of the at least ~~one variation~~ two variations based on position within a sentence or phrase.

Claim 28 (currently amended): A system according to Claim 26, wherein the implicit rules comprise at least one of (A) a number, (B) having no vowels, and (C) constituting at least one of an article, conjunction and preposition.

Claim 29 (currently amended): A system according to Claim 26, wherein the implicit rules comprise normalizing a number of occurrences for each of the at least two variations ~~one variation~~ using at least one of a normalizing function and relative to a source of the each of the at least two variations ~~one variation~~.

Claim 30 (currently amended): A system according to Claim 26, wherein the implicit rules comprise accommodating multiple forms of capitalization for each of the at least two variations ~~one variation~~ by annotating each capitalization form with a frequency count and skipping those of the each of the at least two variations ~~one variation~~ occurring infrequently.

Claim 31 (original): A system according to Claim 17, further comprising:

a hash table maintaining the lexicon.

Claim 32 (original): A system according to Claim 31, further comprising:

at least one record specifying at least one such word as a key into the hash table, and associating at least one such variation within the word set as a preferred capitalization.

Claim 33 (currently amended): A method for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

aggregating each word set, comprising:

identifying at least one word set comprising significant statistics;

selecting at least ~~one~~ two such ~~variation~~ variations within the identified word set having a non-standard capitalization; and

adding the at least ~~one~~ two such ~~variation~~ variations to the lexicon.

Claim 34 (original): A method according to Claim 33, further comprising:

tokenizing the excerpt into the one or more words and one or more punctuation marks.

Claim 35 (original): A method according to Claim 34, further comprising:

splitting hyphenated words into a plurality of the words.

Claim 36 (original): A method according to Claim 33, further comprising:

skipping at least one variation which may be at the start of a sentence.

Claim 37 (original): A method according to Claim 36, wherein the non-standard capitalization comprises the at least one variation occurring in an excerpt having fewer than half of individual letters provided in uppercase.

Claim 38 (currently amended): A method according to Claim 33, further comprising:

normalizing a plurality of the words extracted relative to a source of the unstructured excerpt ~~structured Web content~~.

Claim 39 (currently amended): A method according to Claim 33, wherein the set comprising significant statistics comprises only non-standard capitalization

variations having at least four occurrences of at least one such variation within a word set.

Claim 40 (original): A method according to Claim 33, wherein the non-standard capitalization comprises the at least one variation having any individual letter other than the first individual letter provided in uppercase.

Claim 41 (currently amended): A method according to Claim 33, further comprising:

selecting at least one such variation within the identified word set having a standard capitalization, and adding the at least ~~one~~ two such ~~variation~~ variations to the lexicon.

Claim 42 (original): A method according to Claim 33, further comprising:

applying implicit rules for capitalization; and

skipping each at least one variation subject to at least one such implicit rule.

Claim 43 (original): A method according to Claim 42, wherein the implicit rules comprise skipping at least one variation based on position within a sentence or phrase.

Claim 44 (currently amended): A method according to Claim 42, wherein the implicit rules comprise at least one of (A) a number, (B) having no vowels, and (C) constituting at least one of an article, conjunction and preposition.

Claim 45 (currently amended): A method according to Claim 42, wherein the implicit rules comprise normalizing a number of occurrences for each of the at least two variations ~~one variation~~ using at least one of a normalizing function and relative to a source of each of the at least two variations ~~one variation~~.

Claim 46 (currently amended): A method according to Claim 42, wherein the implicit rules comprise accommodating multiple forms of capitalization for each at least one variation by annotating each capitalization form with a frequency count and skipping those of ~~the~~ each of the at least two variations ~~one variation~~ occurring infrequently.

Claim 47 (original): A method according to Claim 33, further comprising:

maintaining the lexicon structured as a hash table.

Claim 48 (original): A method according to Claim 47, further comprising:

specifying at least one such word as a key into the hash table; and

associating at least one such variation within the word set as a preferred capitalization.

Claim 49 (canceled)

Claim 50 (original): An apparatus for building a lexicon for use in capitalization correction for unstructured excerpts, comprising:

means for assembling a list of word sets from unstructured content, each word set comprising a word and at least one variation on capitalization;

means for aggregating each word set, comprising:

means for identifying each word set comprising significant statistics;

means for selecting at least ~~one~~ two such ~~variation~~ variations within the identified word set having a non-standard capitalization; and

means for adding the at least ~~one~~ two such ~~variation~~ variations to the lexicon.